National Statistical Service

Confidentiality Information Sheet 6

Contributed by Professor David Lawrence (Telethon Institute of Child Health Research)
and Julie Considine (Australian Bureau of Statistics)

> “Agencies and users should work together to promote legislative, regulatory, and dissemination policies and practices that facilitate timely and cost-effective access to data for statistical research and policy analysis but do not permit full and open access by all of the public for any use. If confidentiality issues are not fully addressed in constructive and proactive ways, users face the very real risk of losing access to high quality data.”
>
> Doyle P, Lane JW, Theeuwes JJM, Zayatz LM (2001) Confidentiality, disclosure and data access. *Theory and practical applications for statistical agencies.* North-Holland

Confidentiality

How confidentiality affects research

Confidentialisation techniques are applied to microdata (unit record data where each record represents observations for a person or organisation) to enable them to be made available to analysts and researchers. Without these techniques, access to valuable information for research and analytic purposes would be severely restricted. Although application of confidentialisation techniques generally leads to losses in information availability, when confidentialisation is done well and with knowledge of the key research objectives in mind, the information loss can be minimal. This fact sheet looks at the impact of confidentialisation on information availability for use in research and analysis.

There has always been a debate over the delicate balance between gaining full, unrestricted access to data (for researchers), and the application of confidentiality techniques to protect the privacy of data providers (by data custodians). Selecting the confidentiality techniques to be applied to microdata is a careful balance between fulfilling obligations to protect the identity of individuals and organisations and maximising the information available for statistical and research purposes.

The information required by the research sector is becoming more sophisticated over time. Increases in the use of techniques such as data linkage, data modelling, and data mining mean that researchers' requirements are more detailed and more varied than ever before. As a consequence, there is more pressure on data custodians to provide greater access to microdata through high quality, detailed unit record files.

Users of microdata may be concerned that any reductions or changes made to datasets during the confidentialisation process may affect their ability to undertake analysis or research using the data, or may impact on the results of an analysis. However, generally very few changes need to be made to the dataset, and in most cases these have no impact on statistical analyses.

The goal of confidentialisation is to protect the identity of individual respondents. The main types of data cells affected when confidentialising a dataset are:

- rare events or characteristics;
- unusual data (extreme high or low reported values); and
- low count cross-classification cells.

Generally, statistical analysis is based on observing trends and patterns in data, and most statistical techniques rely on multiple events or individuals with similar characteristics from which to draw inferences. Where there are multiple observations with similar characteristics, the risk of individual identification is low, and the confidentialisation process generally would not result in any change to the data. The low frequency events and unusual values that are targeted in confidentialisation procedures are generally not amenable to statistical analysis.

Types of research

While the data requirements for researchers are highly diverse, research studies can be broadly divided into two main groups—quantitative and non-quantitative. Quantitative studies require sufficient amounts of data for reliable estimates or models; non-quantitative studies may focus on an in-depth analysis of an unusual characteristic or small cohort.

Quantitative research and confidentiality

When undertaking analysis for quantitative based research, analysts require sufficient amounts of data to ensure high quality outputs. If the required variables of interest are available, Confidentialised Unit Record Files (CURFs) should be able to meet the needs of researchers. This is because the confidentiality techniques applied to these files are the same steps that analysts typically apply to the data to ensure robust outputs. That is:

- collapse cells with small counts;
- collapse numeric values into groups;
- recode variables with long tails (extreme high or low reported values); and
- remove, or otherwise treat, outliers in distributions.

For these types of analyses the outputs achieved using CURFs should not differ greatly from the results that would have been achieved using the original non-confidentialised unit record files.

Example: How does confidentiality affect statistical analysis?

Professor David Lawrence, Telethon Institute of Child Health Research (TICHR)

Case study: Cigarette smoking and anxiety disorders

TICHR conducted a study using the ABS' 2007 Survey of Mental Health and Wellbeing expanded Confidentialised Unit Record File (CURF) to investigate the association between smoking and mental health problems. The expanded CURF was analysed through the Remote Access Data Laboratory (RADL). The RADL is a remote analysis environment where a client can submit analysis code to the data custodian, via the internet, and subsequently receive the output without ever having direct access to the unit record level information.

After the project was completed, an evaluation was conducted to determine whether the confidentiality techniques of perturbation and re-coding of variables (high and low data values) applied to the CURF had any effect on the results of the analysis. An ABS officer re-ran the analyses using the Survey of Mental Health and Wellbeing master file (the original file before confidentialisation techniques were applied). The analysis included running several multi-way weighted tables of proportions, and fitting several logistic and proportional hazards regression models. Results from the study were published to three significant digits. None of the figures varied between the CURF and master file analyses at this level of precision.

Table 1 shows estimated hazard ratios for quitting smoking to five significant digits as obtained from performing the same analyses on both the expanded CURF and the original master file. The largest difference was observed at the level of 4 significant digits, and represented 1% of the standard error of the relevant estimate.

The minor changes between the CURF and master file analyses were of neither practical or statistical significance. This result is consistent with the very small level of change to the dataset during the confidentialisation process as described in the User's Guide for the survey CURF.

Summary: When our research on smoking and mental illness was run on the master file, the outputs were not practically different to those achieved using the expanded CURF. Therefore, the confidentiality techniques applied to the master file (perturbation and recoding of variables) did not affect the analyses conducted or the conclusions drawn from them during the project.

Full details of the weighted analyses and models undertaken for this project are described in:

Lawrence D, Considine J, Mitrou F, Zubrick SR (2010) Anxiety disorders and cigarette smoking. Results from the Australian Survey of Mental Health and Wellbeing. *Australian and New Zealand Journal of Psychiatry.* 44: 521-528.

Table 1: Hazard ratios* for smoking cessation, people with anxiety disorders compared with people with no lifetime mental disorder, by type and nature of anxiety disorder

Disorder	Hazard Ratio (from CURF)	Hazard Ratio (from master file)	Standard Error of Hazard Ratio	Difference between hazard ratios as proportion of standard error
No lifetime mental disorder	1.00	1.00	(reference category)	
Anxiety disorders, by type—				
Panic disorder	0.59678	0.59615	0.10335	0.0102
Agoraphobia	0.44936	0.44936	0.08420	0
Social phobia	0.55374	0.55374	0.07973	0
Generalised anxiety disorder	0.33631	0.33631	0.07846	0
Obsessive-compulsive disorder	0.47782	0.47782	0.10674	0
Post-traumatic stress disorder	0.63221	0.63221	0.07865	0
by severity—				
Mild	0.74901	0.74882	0.11495	0.00218
Moderate	0.58942	0.5892	0.08275	0.00447
Severe	0.39196	0.39202	0.06016	0.00249
by use of services—				
Has accessed services	0.54875	0.54858	0.07046	0.00426
No use of services	0.59683	0.59669	0.07113	0.00323
by years since first onset—				
0-2 years	0.73114	0.73089	0.16758	0.00203
2-5 years	0.74544	0.74492	0.20212	0.00346
5-10 years	0.59126	0.59086	0.12845	0.00522
More than 10 years	0.53100	0.53093	0.05969	0.00201

*Hazard ratios measure how often a particular event occurs in one group compared to how often it occurs in another group, over time.

For more information about confidentiality, or to provide feedback on this series,
please email: inquiries@nss.gov.au

Non-quantitative research

Non-quantitative research may include the qualitative investigation of the circumstances surrounding rare events of interest. CURFs do not meet the data requirements for these studies as CURFs are de-identified and subject to confidentialisation techniques. Alternative approaches are needed. This may require the consent of the subjects involved to participate in this type of research.

Example: A qualitative study of children born with a rare birth defect

This type of study may aim to better understand what factors may have led to the occurrence of a rare birth defect, with the ultimate goal of preventing their occurrence or decreasing their incidence. Clearly these studies play an important role in understanding rare cases or phenomena but they require an alternative approach to statistical analysis methods. In this example, administrative microdata may be useful to identify potential participants in a more detailed study. The researcher would require special approval from the custodian to obtain the sensitive data records. This permission could be granted by a body such as a Human Research Ethics Committee constituted under the National Health and Medical Research Council (NHMRC). Further to this, permission and further information may be needed from the parents of the child born with the rare birth defect. This type of research is best conducted by approaching the affected individuals and directly seeking their consent to participate in a research study. Statistical analysis of a microdata file, whether from a survey or an administrative data source, is rarely the most appropriate research design for this type of investigation.

Conclusion

There are generally two types of research: quantitative and non-quantitative. For quantitative research, data custodians are turning to a diverse range of products to meet researchers' needs, including CURFs and data laboratories. Generally speaking, as the small changes that are made to unit record files during the confidentialisation process usually target unusual or extreme values only, the confidentiality techniques used to make the microdata available have little impact on the analyses performed in these types of studies. In many cases the small proportion of information that is lost from the original microdata has no impact on the question being analysed, and often similar steps would be taken by the researcher to prepare the data for the statistical analysis. As a result there will be little to no impact on the analysis using the confidentialised file compared to the original file.

For non-quantitative studies, such as examining unusual characteristics or small cohorts, or linking data sets, the current publicly available data products or methods are not able to meet the researcher's needs. In these cases, researchers are required to complete a more rigorous application process. This may include sign-off from an ethics committee, permission from the data custodian, and possibly, permission from the data provider.

These types of data requests are increasing and researchers and data custodians will need to work together to come up with innovative ways to streamline access to data sets while ensuring that the confidentiality of the data provider is always respected and maintained. This is critical to ensuring the continuation of Australia's high quality data collections.

For more information about confidentiality, or to provide feedback on this series, please email: inquiries@nss.gov.au